# Invenio@HGF – Collaborative repository infrastructure

## Open Repositories 2014 - Helsinki

Alexander Wagner[1], Robert Thiele[2]
**for the Collaboration**

[1]Forschungszentrum Jülich, Zentralbibliothek
[2]DESY Hamburg, Bibliothek

13. June 2014

# Overview

- Partner
- Initial TODO
- Accomplishments
- Lessons learned
- Project group

# Project Partners

| | |
|---|---|
| Deutsches Elektronensynchrotron, Zentralbibliothek | $\approx 2000 + 3000$ |
| Forschungszentrum Jülich, Zentralbibliothek | $\approx 5000 + 1000$ |
| GSI Helmholtzzentrum für Schwerionenforschung, Bibliothek + Base-IT | $\approx 1050$ |
| Deutsches Krebsforschungszentrum, Biblitohek | $\approx 3000$ |
| Maier-Leibniz-Zentrum, Garching | $\approx 300$ |
| RWTH Aachen, Hochschulbibliothek | $\approx 9000$ |
| Museum Zitadelle Jülich | |
| Institut für Experimentelle Kernphysik, Karlsruhe | |

# Project Partners

| | |
|---|---|
| Deutsches Elektronensynchrotron, Zentralbibliothek | $\approx 2000 + 3000$ |
| Forschungszentrum Jülich, Zentralbibliothek | $\approx 5000 + 1000$ |
| GSI Helmholtzzentrum für Schwerionenforschung, Bibliothek + Base-IT | $\approx 1050$ |
| Deutsches Krebsforschungszentrum, Biblitohek | $\approx 3000$ |
| Maier-Leibniz-Zentrum, Garching | $\approx 300$ |
| RWTH Aachen, Hochschulbibliothek | $\approx 9000$ |
| Museum Zitadelle Jülich | |
| Institut für Experimentelle Kernphysik, Karlsruhe | |

## Open for new **Partners**!

# Project Partners

| | |
|---|---|
| Deutsches Elektronensynchrotron, Zentralbibliothek | $\approx 2000 + 3000$ |
| Forschungszentrum Jülich, Zentralbibliothek | $\approx 5000 + 1000$ |
| GSI Helmholtzzentrum für Schwerionenforschung, Bibliothek + Base-IT | $\approx 1050$ |
| Deutsches Krebsforschungszentrum, Biblitohek | $\approx 3000$ |
| Maier-Leibniz-Zentrum, Garching | $\approx 300$ |
| RWTH Aachen, Hochschulbibliothek | $\approx 9000$ |
| Museum Zitadelle Jülich | |
| Institut für Experimentelle Kernphysik, Karlsruhe | |

## Open for new **Partners**!

Serving now $\approx$ **20.000** people (+ visitors)

($\approx$ 260.000 documents + 80.000 Authorities)

# New partner: Deutsches Krebsforschungszentrum


- Largest german bio-medicine research center
- 3000 people:
  1000 researchers inside
- over 90 divisions and groups: research of cancer-producing, risk factors and strategies against cancer
- Replacing of own existing repository system with Invenio@HGF
- Roll out planned for mid of 2015

# Starting out

## Goal

Replace existing systems, at GSI build up from scratch.
User-centric design (users aka scientists)

# Starting out

## Goal

Replace existing systems, at GSI build up from scratch.
User-centric design (users aka scientists)

1. “Learn Invenio” (thanks to CERN ☺)

# Starting out

## Goal

Replace existing systems, at GSI build up from scratch.
User-centric design (users aka scientists)

1. “Learn Invenio” (thanks to CERN ☺)
2. Define wording. . . (different institutions!)

# Starting out

## Goal

Replace existing systems, at GSI build up from scratch.
User-centric design (users aka scientists)

1. “Learn Invenio” (thanks to CERN ☺)
2. Define wording. . . (different institutions!)
3. Build infrastructure: git and friends

# Starting out

## Goal

Replace existing systems, at GSI build up from scratch.
User-centric design (users aka scientists)

1. “Learn Invenio” (thanks to CERN ☺)
2. Define wording. . . (different institutions!)
3. Build infrastructure: git and friends
4. Build more infrastructure: authorities and friends

# Starting out

## Goal

Replace existing systems, at GSI build up from scratch.
User-centric design (users aka scientists)

1. “Learn Invenio” (thanks to CERN ☺)
2. Define wording. . . (different institutions!)
3. Build infrastructure: git and friends
4. Build more infrastructure: authorities and friends
5. Build a **deployment scheme**: InstallInvenio and friends

# Starting out

## Goal

Replace existing systems, at GSI build up from scratch.
User-centric design (users aka scientists)

1. “Learn Invenio” (thanks to CERN ☺)
2. Define wording. . . (different institutions!)
3. Build infrastructure: git and friends
4. Build more infrastructure: authorities and friends
5. Build a **deployment scheme**: InstallInvenio and friends

We need to roll out 10+ instances

# Starting out

## Goal

Replace existing systems, at GSI build up from scratch.
User-centric design (users aka scientists)

1. “Learn Invenio” (thanks to CERN ☺)
2. Define wording. . . (different institutions!)
3. Build infrastructure: git and friends
4. Build more infrastructure: authorities and friends
5. Build a **deployment scheme**: InstallInvenio and friends

We need to roll out 10+ instances
with different data sets

# Starting out

## Goal

Replace existing systems, at GSI build up from scratch.
User-centric design (users aka scientists)

1. “Learn Invenio” (thanks to CERN ☺)
2. Define wording. . . (different institutions!)
3. Build infrastructure: git and friends
4. Build more infrastructure: authorities and friends
5. Build a **deployment scheme**: InstallInvenio and friends

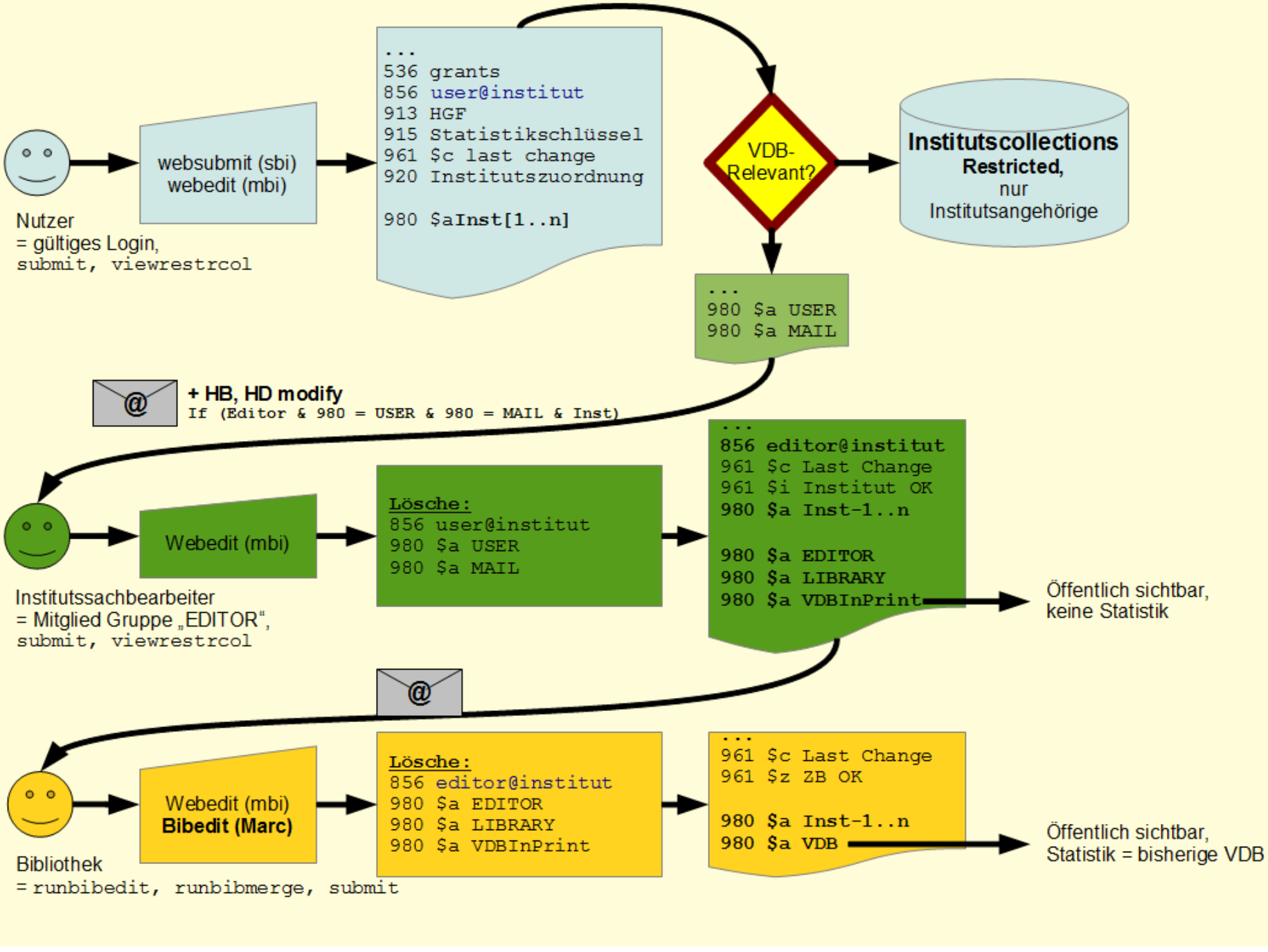
We need to roll out 10+ instances
with different data sets and keep them consistent on code level

# Initial ToDo

- Design the system around web based literature management

# Initial ToDo

- Design the system around web based literature management
- Design a document workflow


Nutzer
= gültiges Login,
submit, viewrestrcol
websubmit (sbi)
webedit (mbi)
...
536 grants
856 user@institut
913 HGF
915 Statistikschlüssel
961 $c last change
920 Institutszuordnung
980 $aInst[1..n]
VDB-Relevant?
Institutscollections
Restricted,
nur
Institutsangehörige
...
980 $a USER
980 $a MAIL
@
+ HB, HD modify
If (Editor & 980 = USER & 980 = MAIL & Inst)
Institutssachbearbeiter
= Mitglied Gruppe „EDITOR“,
submit, viewrestrcol
Webedit (mbi)
Lösche:
856 user@institut
980 $a USER
980 $a MAIL
...
856 editor@institut
961 $c Last Change
961 $i Institut OK
980 $a Inst-1..n
980 $a EDITOR
980 $a LIBRARY
980 $a VDBInPrint
Öffentlich sichtbar,
keine Statistik
@
Bibliothek
= runbibedit, runbibmerge, submit
Webedit (mbi)
Bibedit (Marc)
Lösche:
856 editor@institut
980 $a EDITOR
980 $a LIBRARY
980 $a VDBInPrint
...
961 $c Last Change
961 $z ZB OK
980 $a Inst-1..n
980 $a VDB
Öffentlich sichtbar,
Statistik = bisherige VDB

# Initial ToDo

- Design the system around web based literature management
- Design document workflow (3 steps with privilege escalation)
- Design easy ingestion workflow (websubmit, imports, author disambiguation)

**Preprint**

*Import data* DOI, arXiv, PUBMED...

*ISBN* 978-3-642-12893-6

**Group(s) involved *** T: Theorie-Gruppe

Select or type in name,shortcut (e.g. ATLAS,FS-PE,MKK)

**Beamline/Experiment/Facility (Automatically assigns Grants) *** 514 - Theoretical Particle Physics (POF II: 2010 - 2014)

Select PETRA beamline, HERA, facility machine,...

Grant name / Proposal No. 514 - Theoretical Particle Physics (POF II: 2010 - 2014)

EU project, FS proposal number (e.g. I-20120768)

Report Number DESY 14-075

**Author(s) / Contributor(s) ***

Bonvini, Marco -> Bonvini, Marco (>DESY / T) Corresponding Author

Marzani, Simone [Extern] Author

Start typing lastname and select...

**Title ***

Resummed Higgs cross section at N$^3$LL

Title preview:

Resummed Higgs cross section at $N^3LL$

Journal Type name, issn...; use "" for exact match or fields e.g. title:"nature"

DOI Use Import data for automatic prefill

Volume Issue Pages e.g 47-103

**Publication Year *** 2014 Language Click to select...

Edition 5th ed. Publisher Forschungszentrum Jülich, Verlag

Place of publication Jülich

Title Series

Abstract

We present accurate predictions for the inclusive production of a Higgs boson in proton-proton collisions, via gluon-gluon fusion. Our calculation includes next-to-next-to-leading order (NNLO) corrections in perturbative QCD, as well as the resummation of threshold-enhanced contributions to next-to-next-to-leading logarithmic (N$^3$LL) accuracy, with

# Initial ToDo

- Design the system around web based literature management
- Design document workflow (3 steps with privilege escalation)
- Design easy ingestion workflow (websubmit, imports, author disambiguation)

# Initial ToDo

- Design the system around web based literature management
- Design document workflow (3 steps with privilege escalation)
- Design easy ingestion workflow (websubmit, imports, author disambiguation)

## Finally we wrote some code. . .

# Initial ToDo

- Design the system around web based literature management
- Design document workflow (3 steps with privilege escalation)
- Design easy ingestion workflow (websubmit, imports, author disambiguation)

**Finally we wrote some code. . .**

Every unwritten line is a good line

# Initial ToDo

- Design the system around web based literature management
- Design document workflow (3 steps with privilege escalation)
- Design easy ingestion workflow (websubmit, imports, author disambiguation)

**Finally we wrote some code. . .**

Every unwritten line is a good line, still: $\approx$ 55.000 lines

# Initial ToDo

- Design the system around web based literature management
- Design document workflow (3 steps with privilege escalation)
- Design easy ingestion workflow (websubmit, imports, author disambiguation)

**Finally we wrote some code. . .**

Every unwritten line is a good line, still: $\approx$ 55.000 lines

- Migrate old data (various, proprietary sources)
- Train the inputters and users (secretaries, scientists, librarians)
- Hook up with content management system(s) (visibility!)

# Content management system(s)

# Initial ToDo

- Design a document workflow (3 steps with privilege escalation)
- Establish easy ingestion workflow (websubmit, imports, author disambiguation)

## Finally we wrote some code. . .

Every unwritten line is a good line, still: $\approx$ 55.000 lines

- Migrate old data (various, proprietary sources)
- Train the inputters and users (secretaries, scientists, librarians)
- Hook up with content management system(s) (visibility!)
- Derive necessary reporting (statistics for the Helmholtz Foundation etc.)

# Initial ToDo

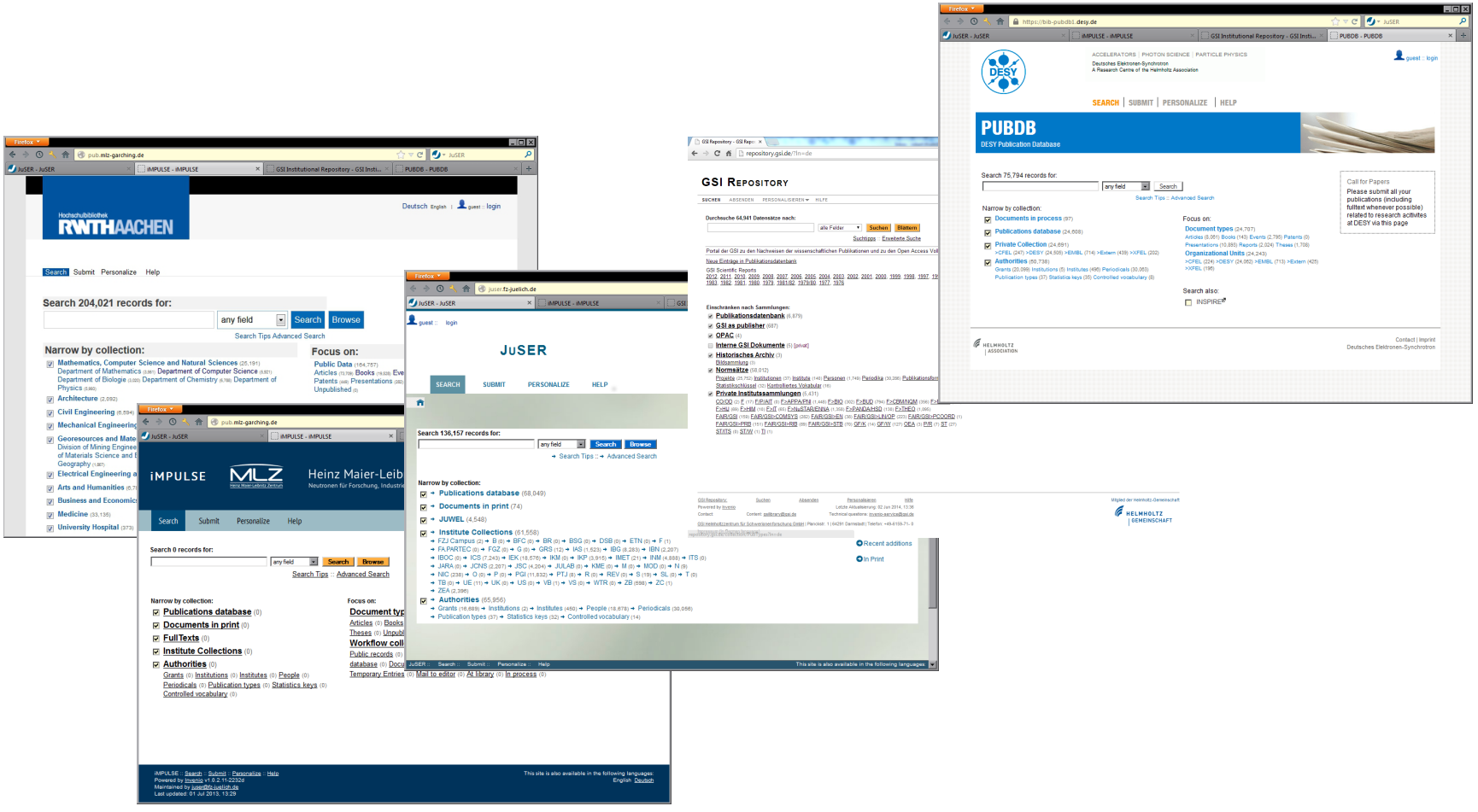
- Design a document workflow (3 steps with privilege escalation)
- Establish easy ingestion workflow (websubmit, imports, author disambiguation)

## Finally we wrote some code. . .

Every unwritten line is a good line, still: $\approx$ 55.000 lines

- Migrate old data (various, proprietary sources)
- Train the inputters and users (secretaries, scientists, librarians)
- Hook up with content management system(s) (visibility!)
- Derive necessary reporting (statistics for the Helmholtz Foundation etc.)
- Get it up and running (First Light: 11/19/2012)

# Accomplishments and status



# Accomplishments and status

- All partners have running systems (roll out works)

# Accomplishments and status

- All partners have running systems (roll out works)
- Almost all partners are online

# Accomplishments and status

- All partners have running systems (roll out works)
- Almost all partners are online
- Rich websubmit (including **repeatable field** handling)
- Importer routines (doi, pmid, arXiv, inspire, ISBN, own recs,. . . in **websubmit**)

# Accomplishments and status

- All partners have running systems (roll out works)
- Almost all partners are online
- Rich websubmit (including **repeatable field** handling)
- Importer routines (doi, pmid, arXiv, inspire, ISBN, own recs,. . . in **websubmit**)
- Authorities

# Accomplishments and status

- All partners have running systems (roll out works)
- Almost all partners are online
- Rich websubmit (including **repeatable field** handling)
- Importer routines (doi, pmid, arXiv, inspire, ISBN, own recs,. . . in **websubmit**)
- Authorities
  - Generate ($\approx$ 80.000 recs)
  - Use (e. g. JSON returns, statistics. . . )
  - Share (MarcXML OAI-PMH)

# Accomplishments and status

- All partners have running systems (roll out works)
- Almost all partners are online
- Rich websubmit (including **repeatable field** handling)
- Importer routines (doi, pmid, arXiv, inspire, ISBN, own recs,. . . in **websubmit**)
- Authorities
  - Generate ($\approx$ 80.000 recs)
  - Use (e. g. JSON returns, statistics. . . )
  - Share (MarcXML OAI-PMH)
- Implement

# Accomplishments and status

- All partners have running systems (roll out works)
- Almost all partners are online
- Rich websubmit (including **repeatable field** handling)
- Importer routines (doi, pmid, arXiv, inspire, ISBN, own recs,. . . in **websubmit**)
- Authorities
  - Generate ($\approx$ 80.000 recs)
  - Use (e. g. JSON returns, statistics. . . )
  - Share (MarcXML OAI-PMH)
- Implement
  - Author identification (ORCiD ready!)
  - Output formats (JSON, BibTEX, EndNote. . . or special formats for our partners)
  - Reporting (publication statistics)
  - Delivery to content management systems

# Statistics tool

## Analyzing collection "VDB" for WEB year 2013

27216 records in collection "VDB"
2527 records relevant for "WEB 2013"

| | | | |
|---|---|---|---|
| 771 | records | JCR listed | (JCR) |
| 6 | records | WOS and not JCR | (WOSnonJCR) |
| 777 | records | Web of Science | (WOS) |
| 777 | records | WOS listed journal OR entry | (WOS |
| 48 | records | Other refereed | (Other |
| 40 | records | Refereed NOT in WoS | (Other |
| 1017 | records | All refereed | (Ref_a |
| 800 | records | Pubmed listed | (Medli |
| 800 | records | Scopus listed | (Scopu |
| 112 | records | DOAJ listed | (DOA |

| | | |
|---|---|---|
| 2 | of type | Diploma Thesis |
| 56 | of type | Dissertation / PhD Thesis |
| 1 | of type | Habil / Postdoctoral Thesis (Non-ger |
| 711 | of type | Internal Report |
| 1103 | of type | Journal Article |
| 2 | of type | Lecture |
| 14 | of type | Master Thesis |
| 8 | of type | Bachelor Thesis |
| 213 | of type | Poster |

### Per program and statistics key

| | JCR listed | WOS and not JCR | Web of Science | WOS listed journal OR entry | Other refereed | Refereed NOT in WoS | All refereed | Pubmed listed | Scopus listed | DOAJ listed |
|---|---|---|---|---|---|---|---|---|---|---|
| 110 - Erneuerbare Energien (POF II: 2010 - 2014) | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 120 - Rationelle Energieumwandlung und -nutzung (POF II: 2010 - 2014) | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 130 - Kernfusion (POF II: 2010 - 2014) | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 140 - Nukleare Sicherheitsforschung (POF II: 2010 - 2014) | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 150 - Technologie, Innovation und Gesellschaft (POF II: 2010 - 2014) | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 210 - Geosystem: Erde im Wandel (POF II: 2009 - 2013) | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 220 - Marine, Ksten und Polare Systeme (POF II: 2009 - 2013) | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 230 - Atmosphre und Klima (POF II: 2009 - 2013) | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 240 - Terrestrische Umwelt (POF II: 2009 - 2013) | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 260 - Technologie, Innovation und Gesellschaft (POF II: 2010 - 2014) | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 310 - Krebsforschung (POF II: 2009 - 2013) | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 320 - Herz-Kreislauf- und Stoffwechselerkrankungen (POF II: 2009 - 2013) | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 330 - Funktion und Dysfunktion des Nervensystems (POF II: 2009 - 2013) | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 340 - Infektion und Immunitt (POF II: 2009 - 2013) | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 350 - Umweltbedingte Strungen der Gesundheit (POF II: 2009 - 2013) | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 360 - Systematische Analyse von multifaktoriellen Erkrankungen (POF II: 2009 - 2013) | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 410 - Supercomputing (POF II: 2010 - 2014) | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 420 - Grundlagen zukünftiger Informationstechnologien (POF II: 2010 - 2014) | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 430 - NANOMIKRO (POF II: 2010 - 2014) | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 440 - Funktionale Werkstoffsysteme (POF II: 2010 - 2014) | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 450 - BioSoft (POF II: 2010 - 2014) | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 460 - BioGrenzflchen (POF II: 2010 - 2014) | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 470 - Technologie, Innovation und Gesellschaft (POF II: 2010 - 2014) | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 510 - Elementarteilchenphysik (POF II: 2010 - 2014) | 285 | 0 | 285 | 285 | 4 | 4 | 341 | 293 | 293 | 49 |

# Tools used

- Workflow

# Tools used

- Workflow
  - Webbaskets (e. g. revision lists)
  - Alerts (e. g. revision lists)
  - Collections (e. g. private for institutes)
  - Webmessage (e. g. correction requests)

# Tools used

- Workflow
  - Webbaskets (e. g. revision lists)
  - Alerts (e. g. revision lists)
  - Collections (e. g. private for institutes)
  - Webmessage (e. g. correction requests)
- Authority records (almost everywhere)

# Tools used

- Workflow
  - Webbaskets (e. g. revision lists)
  - Alerts (e. g. revision lists)
  - Collections (e. g. private for institutes)
  - Webmessage (e. g. correction requests)
- Authority records (almost everywhere)
- OAI-PMH (authority exchange)

# Tools used

- Workflow
  - Webbaskets (e. g. revision lists)
  - Alerts (e. g. revision lists)
  - Collections (e. g. private for institutes)
  - Webmessage (e. g. correction requests)
- Authority records (almost everywhere)
- OAI-PMH (authority exchange)
- High-level API (setup: e. g. collections, roles, groups, baskets. . . ; **no db-dump** sharing)

# Tools used

- Workflow
  - Webbaskets (e. g. revision lists)
  - Alerts (e. g. revision lists)
  - Collections (e. g. private for institutes)
  - Webmessage (e. g. correction requests)
- Authority records (almost everywhere)
- OAI-PMH (authority exchange)
- High-level API (setup: e. g. collections, roles, groups, baskets. . . ; **no db-dump** sharing)
- jQuery/jQueryUI (websubmit)

# Tools used

- Workflow
  - Webbaskets (e. g. revision lists)
  - Alerts (e. g. revision lists)
  - Collections (e. g. private for institutes)
  - Webmessage (e. g. correction requests)
- Authority records (almost everywhere)
- OAI-PMH (authority exchange)
- High-level API (setup: e. g. collections, roles, groups, baskets. . . ; **no db-dump** sharing)
- jQuery/jQueryUI (websubmit)
- intbitsets (e. g. statistics)

# Lessons learned / Next steps

- CERN is way to fast to keep up with

# Lessons learned / Next steps

- CERN is way to fast to keep up with
- Never use Dublin Core again (complex migration, to few data fields. . . )

# Lessons learned / Next steps

- CERN is way to fast to keep up with
- Never use Dublin Core again (complex migration, to few data fields. . . )
- All libraries are the same ☺

# Lessons learned / Next steps

- CERN is way to fast to keep up with
- Never use Dublin Core again (complex migration, to few data fields. . . )
- All libraries are the same ☺

- Upgrade to 1.1.x:
  - get OAI-Server fixed, no hanging bibsched, etc.
  - testing and bugfixing on our test systems
  - roll out update in July by our partners. . .

# Lessons learned / Next steps

- CERN is way to fast to keep up with
- Never use Dublin Core again (complex migration, to few data fields...)
- All libraries are the same ☺

- Upgrade to 1.1.x:
  - get OAI-Server fixed, no hanging bibsched, etc.
  - testing and bugfixing on our test systems
  - roll out update in July by our partners...

**However...**

In our use case switching of the base system is non-trivial

(Remember: 10+ instances...)

# Lessons learned / Next steps

- CERN is way to fast to keep up with
- Never use Dublin Core again (complex migration, to few data fields...)
- All libraries are the same ☺

- Upgrade to 1.1.x:
  - get OAI-Server fixed, no hanging bibsched, etc.
  - testing and bugfixing on our test systems
  - roll out update in July by our partners...

**However...**

In our use case switching of the base system is non-trivial

(Remember: 10+ instances...)

- Open up for new partners
- Clean up our code and give it back → moving to github

# Contributors

- *Martin Köhler*[a]
- *Robert Thiele*[a]
- Katrin Große[b]
- *Stefan Hesselbach*[c]
- Bernhard Mittermaier[d]
- Anna Fründ[d]
- Heike Lexis[d]
- *Cornelia Plott*[d]
- Christoph Holzke[d]
- *Alexander Wagner*[d]
- Dagmar Sitek[e]
- *Gudrun Friedburg*[e]
- Jürgen Neuhaus[f]
- Connie Hesse[f]
- *Björn Pedersen*[f]
- Ulrike Eich[g]
- *Louai Barake*[g]
- *Abdoulaye Diallo*[g]
- Roland Rappmann[g]
- Dominik Schmitz[g]
- Edmund Wollgarten[g]

[a] DESY Library and Documentation; [b] GSI Library and Documentation; [c] GSI Base-IT;

[d] Forschungszentrum Jülich, Zentralbibliothek; [e] DKFZ Heidelberg; [f] MLZ, Garching; [g] RWTH Aachen, Hochschulbibliothek

# Further “reading”

- Invenio @ HGF - Technical background

  Talk at Invenio Developer Forum

- Collaborative tools for an institutional repository

  Talk at Helmholtz OA Workshop

- JuSER – Publications Database

  Introductory course at Jülich

- JuSER - Autorenhandling

  Talk at HGF-ORCiD Meeting, Berlin (in german)

- Invenio @HGF – status and perspectives

  Talk at 2nd Invenio User Group Workshop, [sic!], Jülich, Germany

- The Helmholtz INVENIO Repository Project

  Talk at SACITiL-2014, Kolkata, India

# Thanks!

Robert Thiele
DESY-Bibliothek

Subject Specialist for Photon Science

Tel.: +49-40-8998-1927
robert.thiele@desy.de

This document is available as

DESY-2014-02793 or FZJ-2014-02848

*Typeset by pdfLaTeX*